caBIG Meeting Record

Integrative Cancer Research/Genome Annotation SIG Teleconference
3:00 – 4:00 PM ET
December 2nd, 2004

Teleconference Information: (Phone) (Passcode) (Moderator)	 Toll Free Number: 800-593-0616 Passcode: 8511747 Moderator: Craig Street/Juli Klemm
Net-Centra Information: (URL) (Meeting ID)	
Key Decisions and/or Outcomes: *(copy & paste from agenda)*	▪ An update on the GoMiner project was presented. ▪ An update was provided on the Gene CDE discussion with Denise Warzel from the VCDE group. ▪ A modeling strategy for gene and gene product identifiers that provides a means for data integration within the caBIG was presented and discussed.
Executive Summary:	This teleconference started with a status update on the GoMiner project. This was followed by a discussion regarding the modeling strategy for gene and gene product identifiers for data integration within caBIG. Lastly, there was a brief overview of potential external vocabularies and ontologies to be considered for adoption.
Action Items:	▪ Craig Street/Juli Klemm/Rakesh Nagarajan/Vishal Nayak will field questions on the modeling strategy discussed for gene and gene product identifiers and report them to the group. ▪ Genome Annotation SIG participants will provide comments/feedback/suggestions on the modeling strategy for gene and gene identifiers. ▪ Craig Street will send the meeting minutes on December 3rd 2004 to the Genome Annotation SIG. ▪ Genome Annotation SIG participants will review the meeting minutes and send comments back to Craig Street in 2 days.

caBIG cancer Biomedical Informatics Grid

caBIG Meeting Record

Next Steps:	The next teleconference will focus on feedback from the Genome Annotation SIG members on the modeling strategy for gene identifiers presented in this meeting and attempt to come up with a final document/whitepaper. Members should also consider, in advance of the next meeting, what other Common Data Elements should be curated by this SIG (e.g., Gene Products (mRNA, Protein), Polymorphisms, Pathways. Next meeting will further discussions on the external vocabularies and ontologies that may be incorporated into caBIG.
Attendees:	The following were present: • Juli Klemm (BAH) • Rakesh Nagarajan (Washington University) • Baris Suzek (Georgetown) • Mary McAdams (INS) • David Kane (SRA) • Rob Sfeir (SRA) • Kutbuddin Doctor (Burnham Institute) • Terry Braun (University of Iowa, Holden) • Ed Frank (Argonne National Lab) • Brian Gilman (Cold Spring Harbor/ Panther Informatics) • Mike McCormick (Fred Hutchinson) • Lee Davis (Fred Hutchinson) • Chris Abajian (Fred Hutchinson) • Harold Riethman (Wistar) • Vishal Nayak (Penn) • Craig Street (Penn) • Xiaopeng Bian (NIH) • Jong Dang (Alpha Gamma Technologies) • Zhangvhi Hu (Georgetown)
Detailed Meeting Notes: *(copy & paste from agenda and place your meeting notes under each agenda item)*	**3:00 - 3:05** **Roll-call, open meeting, review meeting goals (Craig Street)** **3:05 - 3:15** **Update on GoMiner project (Rob Sfeir)** ▪ The requirement and use case documentation has been prepared. ▪ A number of SOAP prototypes have been written to acquire data on BioCarta Maps from caBIO. ▪ A brownbag lunch was held to conduct discussions on how to connect to caBIG with GoMiner. ▪ The project members have been able to establish connection

to the CVS repository.

3:15 - 3:45

Update on Gene CDE discussion with Denise Warzel and the VCDE group/ Presentation of a modeling strategy for gene and gene product identifiers to enable data integration within caBIG (Craig Street/ Rakesh Nagarajan/Juli Klemm/Vishal Nayak)

- Rakesh Nagarajan presented his approach to modeling gene and gene product identifiers, which will enable integration of disparate sources of data within caBIG.
- Develop a general identifier class which contains a list of pre-approved gene/gene product identifiers.
- This approach provides the flexibility to individual developers to develop their own biological classes (e.g. gene/protein/mRNA class)
- These classes are constrained in having to reference the identifier class with at least one of the attributes NOT NULL.
- A join can be performed on classes having a common identifier.
- Denise Warzel had pointed out that the UML Loader for the current caDSR model does not incorporate a method to force at least one of the attributes of the identifier class to be NOT NULL.
- This identifier class could pave the way for a mapping service in the future.
- The attribute list for the identifier class is extensible and may b expanded to include more identifiers, if approved.
- Ed Frank asked if hard-wiring a particular class to the identifier class could create compile or runtime issues. Rakesh pointed out that these issues would not arise (since the CDEs are not compiled) and that designing the identifier class in this manner would not allow for the enforcing of the one attribute being not null constraint. Ed Frank also mentioned that this hard-wiring approach would be inflexible when it comes to incorporating changes. Rakesh replied that the identifier class will be versioned as changes are made and thus be flexible to change.
- Ed Frank wanted to know that if a particular developer wanted to call a particular service, how would he/she know what kind

of identifier is acceptable to that service. Rakesh replied that the mapping service (when deployed) would fill out the identifier object with all other values when it is passed to that service. Ed Frank then questioned how this issue of knowing which identifier a particular service would require would be handled in the pre-mapping service availability stages. Also, to reduce the load on the mapping service, he wanted to know if an approach could be identified where the mapping service only fills out the identifier values that a particular service requires before passing on the identifier object. He then suggested having a sketch of an interaction diagram showing a caller, a service, and the mapper making use of a proposed interface utilizing identifiers to achieve a call.

- Ed Frank asked if the type of identifier (LocusLink, UniProt, etc.) could be enforced in the value domain while having a general data element concept called Gene Identifier. However, since the value domain has to be curated manually, this presents difficulties.
- Terry Braun stated an example of the need for a mapping service (and how this particular approach is a step in that direction) in which users in his lab need to find the curated RefSeq number from the GenBank accession number and then with the RefSeq number, use an Ensembl module to find the corresponding Ensembl ID and the associated identifiers. Terry Braun inquired about needs for data models beyond an identifier class. He was unclear about how to reuse the CDEs to create data models. He wanted to know whether they should evaluate existing data models and extend them (if they are close to what they are looking for) or create their own data models and load CDEs into caDSR.
- Brian Gilman asked if this is any different from GO (Gene Ontology). Craig Street said that GO is not meant for mapping among different identifiers, but is meant to map molecular function, biochemical activity, and cellular location to genes; thus the GO hierarchy does not deal with gene identification. He informed that GeneLynx could be housed locally to facilitate our mapping service.
- Craig Street mentioned that this identifier class could have a primary identifier that could function as a caBIG identifier. The pros and cons of having and maintaining an additional identifier were weighed. There was not a consensus as to whether this is an identifier that would be used internally or allow outside entities to reference.

- Baris Suzek wanted to know if joins can be performed only if the genes have the same type of identifier (regardless of whether it is the same gene). Rakesh Nagarajan replied in the affirmative as the mapping service is yet to be developed.
- Juli Klemm suggested that every participant provide input and direct any questions they have about this method to Craig Street or the group as a whole and work towards developing a consensus by the next SIG meeting in January.
- Juli Klemm also replied to a question by Terry Braun that the Gene symbol is a very ambiguous concept and is not to be included in the identifier class. She also mentioned that since the NCIB is moving away from the LocusLink ID to the Entrez Gene ID, these changes will have to be incorporated in future versions of the Identifier class.

3:45 - 3:55

Discussion on what external vocabularies and ontologies should be adopted into caBIG
(Discussion on what additional CDEs should be curated by the Genome Annotation SIG was postponed due to time constraints)

- Craig Street talked about OBO (Open Biological Ontologies). This is a collection of open source, well-structured, and orthogonal ontologies that can be freely adopted by biological researchers.
- Baris Suzek talked about UniProt which is a keyword controlled vocabulary. It has about 1000 keywords which are a combination of SwissProt and PIR keywords. Baris also mentioned that this ontology is hierarchical (protein families) and mapped to GO. He also mentioned that the best source for post-translational modifications is RESID. There are 339 modifications listed in the RESID database.
- Harold Riethman provided an introduction to ENCODE (*ENC*yclopedia *O*f *D*NA *E*lements) which is an NHGRI funded project. The purpose of this project is to identify all the functional sequences in the human genome. It systematically targets particular regions and is meant for high throughput annotation. This can be done using techniques such as comparative sequence analysis and large scale chipping. It is an open ended project.
- Juli Klemm talked about getting ontologies into the EVS (Enterprise Vocabulary Services). The EVS provides APIs to

	directly use the protein ontologies. The controlled vocabularies can be used to enforce constraints on value domains. **3:55 - 4:00** **Set agenda for next meeting (All)** ▪ Digest the information presented at this meeting and be prepared to come to a consensus at the January meeting so that a paper may be distributed to the other SIGs in ICR. ▪ Identify other CDEs to be curated by the Genome Annotation SIG. ▪ Further discuss potential external vocabularies and ontologies to be incorporated into caBIG.
Other *(copy & paste from agenda)*	